Dear GMD Editor(s),

Thank you for handling the manuscript. We have completed a round of revisions. Most of the comments are asking for clarification, except for the first point of RC1 who asks about the novelty of the work. We must say that although the model structure has not changed from the last paper, there are substantial uncertain questions regarding effectiveness, generalizability, and strengths and limitations when differentiable models and LSTM are applied to the global scale. We gained lots of insights from this paper, including what state-of-the-art prediction accuracy these models can achieve at the global scale by learning from large data; how different models behave in global catchments across varied climate groups and how their performance patterns can be related to geographic characteristics and hydrologic processes; what physical representations need to be improved towards developing stronger global models; and which modeling approach may be more appropriate to serve as next-generation global models to characterize the hydrologic processes in ungauged regions across the globe.

None of these findings would have been clear unless we attempted to apply these models at the global scale. Hence we think this work has suitable value for the community. This paper was classified as a Model Evaluation paper which we think appropriately captures the type of the study. It also serves as an important benchmark and model diagnosis for future global hydrologic modeling.

RC1:

This study presents a simulation using a differentiable model at 3753 basins globally. It represents an advance in the field and is thus worthy of being published somewhere. However, the used model and the design of the experiments are almost the same as those published in the authors' previous papers. Readers of GMD would expect more progress in those aspects. Thus, I suggest a major revision.

We thank you for your comments. We would like to note, although the main model structure has not been changed, the dataset, training and related experiments changed drastically. We expanded the dataset from only the USA to the whole globe with much larger numbers of basins with diverse climate and geographic attributes, which represents a substantial amount of work. Enormous efforts went into cleaning the data, improving model efficiency, training and testing, and interpreting the results. It was highly uncertain if the model qualities shown at CONUS scale would carry to such a large dataset at global scale. Many of the hydrologic insights are new. The novel experiment cross-continent predictions examined if differentiable models are strong candidates for solving the practical issue of prediction in ungauged regions at global scale. Plus, we do want to ensure some consistency between the evaluation across datasets so we can understand how the same model structure behaves on different datasets. We revised the manuscript to better highlight our novel research questions, as shown below. Please also see our responses to other comments to address this point.

This paragraph is modified to explain the progress of this study to our previous studies,

"We desire efficient regionalized models that maximally leverage available information and provide accurate predictions to diverse basins across different climate groups and geographic characteristics in the world. We also want the models to perform decently even in data-sparse regions, showing competitive extrapolation ability, given that many large regions such as in Africa and Asia lack publicly available streamflow data. DL and differentiable models seem plausible candidates for such simulations. Nevertheless, previous studies on DL and physics-informed differentiable models mainly focus on continental or smaller scales, with a relatively homogeneous forcing dataset --- it is unclear if their observed strengths, e.g., high performance and strong generalization ability, can carry over to global scales, where the climate is much more diverse and datasets differ widely in their biases and uncertainty characteristics. In particular, we want to thoroughly examine how well these models can leverage information learned in data-rich continents to characterize the hydrologic processes in ungauged regions across the world. Meanwhile, DL models also show favorable scaling relationships (or data synergy) where more data leads to more robust models (Fang et al., 2022). Thus, training on a larger dataset may provide additional benefits." lines 97-108

We also added these statements and discussions to clarify this,

"We first investigate what prediction accuracy can be achieved by different models at global scale by learning from a large and diverse dataset. We then relate the global spatial patterns of model performance to geographic characteristics and hydrologic processes to identify model structural deficiencies and gain hydrologic insights. Finally, we provide evidence indicating which type of model may be more appropriate for next-generation global modeling by rigorously examining their generalizability to ungauged regions across the world." lines 117-121

"This study builds a benchmark and a basis for model selection and diagnosis for the next-generation global hydrologic modeling, which previously did not learn from such large observations. With rigorous tests at global scale, this study proves that differentiable models are strong candidates as global water models. With powerful spatial generalization ability, they can be applied to characterizing the hydrologic processes in ungauged regions by leveraging learned information in data-rich continents. Differentiable models in this study have already learned the generic and robust relationships between geographic attributes and physical model parameters from thousands of global catchments. Therefore, these models can be easily applied towards providing seamless global hydrologic modeling with parameters directly generated from worldwide geographic attributes. Future work can use such models to produce global hydrologic fluxes while enhancing some process representations in extremely arid, glaciated, or heavily human-influenced basins." lines 427-435

GMD specifically has a model evaluation paper category, and there are many papers on GMD which evaluate the same model under different scenarios. In fact, we can probably say that it is a minority case to be changing the model structure significantly from paper to paper. As such, we

think the paper fits GMD very well, and will be a valuable contribution to the model evaluation literature.

Major comments

- The authors should introduce more details of the experimental design, such as the metrics used in the training, how many experiments (temporal generalization, PUB, and PUR; correct me if I am wrong), and the purpose of the experiments. Some details may have been presented somewhere else. I find this manuscript difficult to follow without reading the authors' previous publications.

Thanks for the comment. We have revised the manuscript to give some more background on earlier work, and added details for metrics used, experimental design, etc.

"As in Feng et al., (2022), we employed the loss function based on root-mean-square error (RMSE) with two weighted parts. The first part calculates RMSE directly on the simulated and observed discharge, while the second part calculates RMSE on the transformed discharge records to improve low flow representations." lines 179-182

Section 2.4 explains all the experiments and and we modified it as,

"We ran one temporal and two spatial generalization experiments to evaluate the performance of different regionalized models. For the temporal generalization experiment, the models were trained for the period of 2000 to 2016 on all global basins, and tested for the period of 1980 to 1997. Without spatially holding out any basin during training, this experiment aimed at evaluating the model's generalizability in the time dimension by testing prediction ability on the same basins but in a different time period from the training data. The other two spatial generalization experiments served as the true litmus tests for evaluating the effectiveness of regionalization schemes, i.e., how well the model can be applied to basins that have never been seen during training. The first spatial generalization experiment was a traditional "prediction in ungauged basins" (PUB) problem, where we randomly divided the whole global basin set into 10 folds (groups) and performed cross-validation across these folds to obtain spatial out-of-sample predictions for all basins (training on 9 of the folds with the 10th fold held out and testing on the 10th, then rotating such that each fold is used for testing once). The second spatial generalization experiment, which we refer to as cross-continent "prediction in ungauged regions" (PUR), was more challenging. In this experiment, we assumed that all the basins in certain continents are ungauged and excluded from the training dataset, trained a regionalized model in other data-rich continents, and then tested the trained model to make predictions in the ungauged continents. With random hold-out, an ungauged test basin in the first spatial generalization experiment always has training gauges surrounding it. Therefore, the first PUB experiment can be interpreted as spatial interpolation. The second spatial experiment (cross-continent PUR) holds out all the basins in one continent as testing targets, and thus is the

much harder test of spatial extrapolation." lines 202-218

We also clarified that *"All the reported performance metrics in this study are from model evaluation on the testing dataset, which is not seen by the model during the training process."* lines 225-227

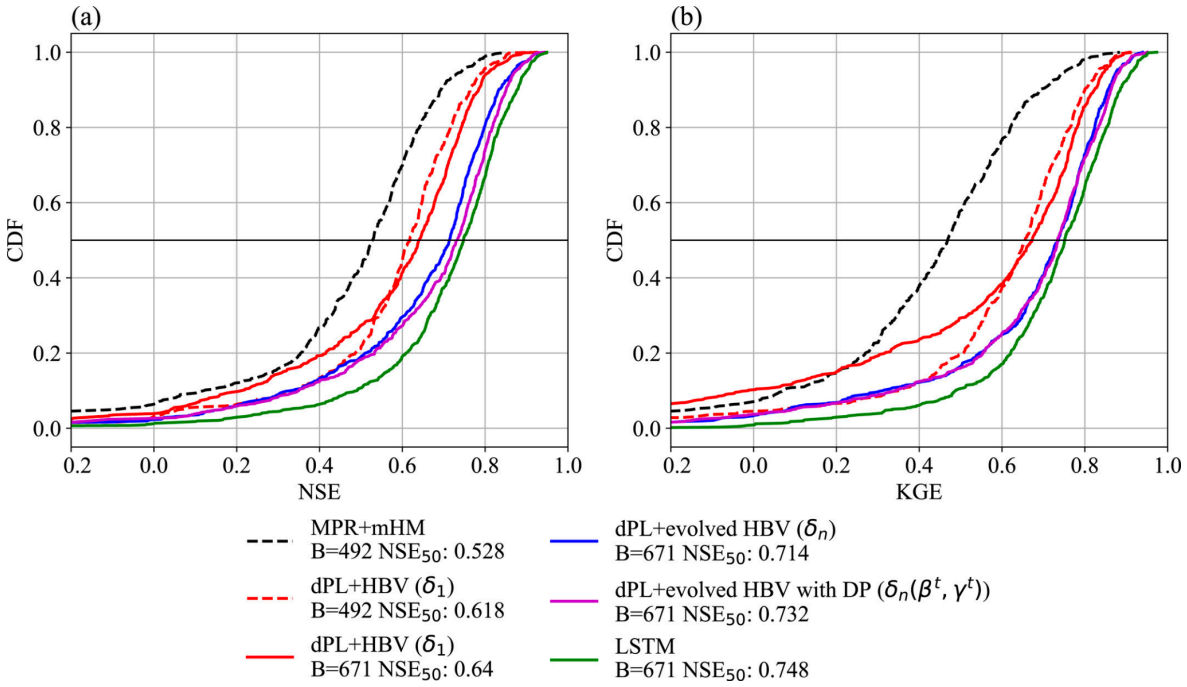
- L124: what are the criteria for the selection? Can you describe the erroneous cases? Do the erroneous cases include the data processing error described in L350?

There are some data records that just seem totally unreasonable, for example, total discharge records are way larger than precipitation, yet unable to be linked to some simple unit conversion errors. Some basins also have discharge values that show differences of several magnitudes for two time periods. Line 124 and Line 350 in the original manuscript refer to the same group of gages with records showing these phenomena. We excluded those basins with significant issues but kept those gages showing moderate issues. We now clarified this statement as *"We excluded some basins with potential erroneous discharge records such as showing unreasonable magnitude way larger than precipitation or dramatic differences between two time intervals, by manually performing visual screening"*. Lines 132-134

- L306-L317: Can you discuss more about comparing the traditional regionalization method and PUR? The PUR regionalization method can utilize a large number of observations to calibrate/train the differentiable model, whereas the traditional method can only use very limited samples. In other words, PUR may have a much higher chance of finding the optimal parameters than the traditional method.

Here we compare traditional parameter regionalization methods with differentiable modeling. PUR and PUB, as spatial generalization tests, are the experiments to test the performance of different regionalization methods. In this study, differentiable models largely outperform the traditional method built in Beck et al., 2020 that represents the previous best spatial generalization results at global scale. It has also been established in past work that differentiable models have better performance than another traditional regionalization scheme, MPR (see the large gap between MPR+mHM and all the different models in the below figure from Feng et al., 2022). These traditional methods actually have a similar concept to differentiable modeling, which aims at building the relations between geographic attributes and model parameters, but the traditional methods only use linear functions because the model is not differentiable and they cannot optimize a large number of weights (as in neural networks) like the differentiable models can. Differentiable models instead don't assume any linear transfer function forms for the relations, and automatically learn the complicated relations by the embedded neural networks from big data. Hence what the reviewer said is true, but we would modify it a bit as "differentiable models have a much higher chance of finding the most high-performing, robust, and transferable parameters compared to the traditional method." We added these statements to incorporate the discussion,

"Differentiable modeling does not rely on strong assumptions of the functional form for the parameter transfer function. It leverages the powerful ability of neural networks to represent complicated functions, and automatically learns robust and generalizable relationships between geographic attributes and physical model parameters from large data. Therefore, we can expect significant performance advantages from differentiable modeling compared to traditional methods relying on linear transfer functions." lines 336-340



Minor comments

- Title: the phrase, global hydrologic simulations, is misleading. The simulations are conducted at 3753 basins across the globe. It represents a concept different from the "global hydrologic simulations."

We revised the title to *"Deep dive into hydrologic simulations at global scale"* to avoid confusion. However, it's worth emphasizing that differentiable models shown in this study have already learned the relationships between geographic attributes and model parameters at global scale by learning from thousands of catchments. This means the models can be easily applied for seamless global modeling with parameters directly generated everywhere from geographic attributes using the trained models, which will be the subsequent work following this study. We can imagine that for global seamless simulations, the largest barrier would be generating reliable parameter fields, which is largely solved in this study. We also added these statements to further clarify it,

"Differentiable models in this study have already learned the generalizable and robust relationships between geographic attributes and physical model parameters from thousands of global catchments. Therefore, these models can be easily applied towards providing seamless global hydrologic modeling with parameters directly generated from worldwide geographic attributes. Future work can use such models to produce global hydrologic fluxes while enhancing some process representations in extremely arid, glaciated, or heavily human-influenced basins." Lines 431-435

- L127: is the classification from Beck et al., 2020b?

Yes. We modified the statement as *"These basins had been classified into five Köppen-Geiger climate classes in Beck et al., (2020b), including tropical (489 basins), arid (109 basins), temperate (1423 basins), cold (1593 basins), and polar (139 basins), as shown in Figure 1"*. Lines 136-138

- L26 & L212: How is the subset of 1675 basins selected? What is the objective of the selection?

The selection criteria was long-term streamflow data availability. Here, "long-term" means that these basins have at least 15 years' worth of records (not necessarily continuous) in the training time period and 5 years' worth of records in the testing period for model evaluation. We have modified the following statement in the main text,
"For a subset of 1675 basins with long-term records (at least 15 years' worth of streamflow data available in the training period and 5 years' worth of data available in the testing period, though not necessarily continuous), LSTM even reached a KGE of 0.78." Lines 234-236

- L223: can you describe more about the structural issues? Why does the explicit solution scheme introduce numerical errors here?

Here we mentioned numerical errors introduced by the explicit scheme. There are also issues related to how groundwater storage cannot feed back to the upper layers. We modified the original statement as *"the errors with peak flow could also be partly due to some numerical and structural issues with the differentiable models, e.g., numerical errors introduced by the explicit and sequential solution scheme of HBV with excessive use of threshold functions that lead to different results when the sequence changes, and structure limitations, e.g., deeper groundwater storage cannot feed back to the upper layers."* Lines 248-251

- L291-L293: please rewrite the sentence. It is difficult to read.

We revised the statement to *"The sensitivity of model performance to missing processes in the differentiable models is both good and bad news. It is good news because this means we can identify suitable or insufficient process representations by learning from data..."* lines 318-320

- L398, "the underrepresentation of the processes...": this conclusion is too general. The difficulty of representing arid/polar/anthropogenic processes is known before reading this paper. The conclusion should be specific.

There are many structural deficiencies with HBV, but it seems the neural network based parameterization approach in differentiable modeling can make up for some of the deficiencies. Here we are saying that in these two cases, even the NN-based parameterization cannot make up for these missing processes. We largely clarified these statements as the following:

"For the polar group, the differentiable model performed significantly worse than the LSTM. Without any physical constraints, LSTM shows strong power in simulating storage (snow and glacier) dominated processes, while differentiable models are limited by the structure of their physical backbone model, which in this case does not simulate multiyear ice buildup and melt. Another limitation could be soil sealing processes in extremely arid regions. These regional performance comparisons thus reveal some deficiencies of the physical backbone in δHBV that cannot be mitigated even by advanced neural network-based parameterization." Lines 448-454

CEC1:

Dear authors,

I have checked the "Code and Data Availability" section in your manuscript. I would like to point out that the sentence saying that an updated version of the code and data will be available upon acceptance is misleading. Obviously, if your manuscript is accepted for publication, you have to publish it with the most updated code and data, and we do not accept "upon acceptance" statements. Right now, your statement seems to imply that you have not published your code and data in advance, however, they are included in the Zenodo repository. I would like to ask you to modify the sentence to avoid confusion.

Also, please clarify which is the PyTorch version that you have used for your work.

Regards,

Juan A. Añel

Geosci. Model Dev. Executive Editor

Thanks for the notice. The models themselves were indeed published previously. They are applied to a new (larger) global dataset to yield new insights, since they were not previously applied on the global scale. We have modified it in the revised manuscript along with clarifying the PyTorch version as shown below.

"...was implemented in a DL platform (PyTorch 1.0.1 was used for the original development and the model has also shown good compatibility with more recent PyTorch versions, (Paszke et al., 2017))..." Lines 176-178

"The source codes for the differentiable hydrologic models can be accessed at https://doi.org/10.5281/zenodo.7091334, and this study evaluates these models at global scale. " Lines 473-475

RC2:

Overall, the manuscript is well-written with a clear research objective, innovative hybrid models, and solid results. The study compared the performance of the commonly used LSTM model with two differentiable hydrological models (static and dynamic parameters) with a temporal generalization experiment and conducted a comparative analysis for traditional "prediction in ungauged basins" problem. The model evaluation results provide valuable insights into improving the mechanism of the hydrologic models, confirming the strong localization and extrapolation capabilities of differentiable models. Below are some key comments and concerns:

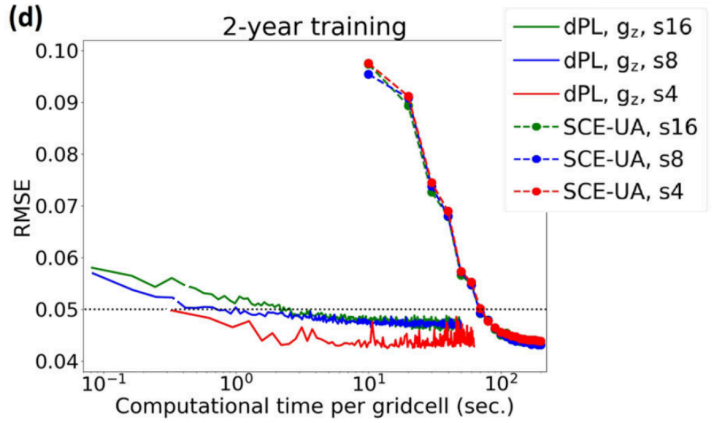
Thank you very much for the evaluation. Please see our responses to each of the comments below.

L161-166: From my understanding, the original parameter calibration process of the HBV model has been replaced by the parameter calibration process of the *gA* neural network. If this is the case, it is still necessary to run the HBV model. How does this approach compare to the traditional hydrological model calibration methods in terms of modeling speed? Has it resulted in time and labor savings in the modeling process?

Good question. The HBV model has been reimplemented on the PyTorch platform, gaining the ability to track gradients and run in high concurrency on GPUs. To accommodate high concurrency, we had to reimplement the code in a parallel way. We further require that a generic parameterization network is trained over all the thousands of sites rather than being calibrated site-by-site, learning a robust mapping relationship. This differentiable modeling framework resulted in orders of magnitude of computational savings, and we'd argue labor saving as well, but it depends on the economy of scale. Below we provide some more background. We realize that many people have not read our previous paper (Tsai et al., 2021), and we include some pieces of the discussions below in the revised manuscript.

There is an obvious time saving in parallelism. During training, we run models in parallel over a "minibatch", which means integrating data from 256 basins at a time, before we calculate the loss and update the weights. This efficient parallelism is due to easy GPU concurrency via PyTorch.

There is a less obvious economy of scale, which is that the training and the learned knowledge of neural networks are shared by all sites. This means the more sites participating in the training, the more efficient it becomes (see the figures below, from Tsai et al., 2021). Hence, if you run differentiable model training for one of a few sites, it won't be as efficient as traditional optimization. If you run it for many more sites, it will become orders of magnitudes more efficient. The number we cite in Tsai et al., 2021 is that, for the same calibration job over the USA, a traditional evolutionary algorithm would need a 100-processor cluster to run 2-3 days, whereas our differentiable parameter learning method only needs 1 single GPU for one hour for a surrogate model. Similarly, only several hours are required for the differentiable HBV with 16 components (the original HBV does not have this subbasin-scale heterogeneity).



More importantly, this differentiable model can achieve several things for which the computational cost would be impossible if using a traditional calibration approach. We have flexibly evolved the structure of the original HBV in the differentiable modeling approach by using multi-component computation and dynamic parameters, as stated in line 177 of the original manuscript. The first version increases HBV parameters from 12 to 12*16=192 with 16 components and the second version has a different parameter value at each time step, varying with the meteorological forcings. It's unrealistic to use traditional calibration to achieve the optimization for these models. However, all these static and dynamic HBV parameters can be automatically and efficiently learned with differentiable modeling.

We argue it actually saves human effort as well, because differentiable modeling enables robust parameter regionalization, which means it optimizes and learns the relations between catchment geographic attributes and model parameters. We only need to set up and train the model once and then it can generate parameter fields across the world and be applied to prediction in ungauged regions as shown in this manuscript. We were never able to do this easily in the past.

We added these discussions to the main text,

"Some central differences exist between the differentiable modeling approach and the traditional calibration approach: First, we always attempt to train differentiable models over a large

collection of sites, improving the robustness and efficiency of learning. Second, we reimplemented the model onto PyTorch in a parallel fashion, and were thus able to leverage the high concurrency and computing power offered by modern GPUs. Third, the commonalities between sites and the accumulation of knowledge in the neural network further improves the efficiency of training compared to traditional training done individually for each site, as the knowledge learned from one batch is inherited when the neural network is trained on the next batch. Fourth, differentiable models can flexibly evolve the structure of the physical backbone. The two types of differentiable models used in this study have used multiple parallel components per basin to represent subbasin-scale heterogeneity. For the models with dynamic parameterization, two parameter values vary at each daily time step as a function of the meteorological forcings. It seems unrealistic to use traditional parameter calibration to optimize these models with evolved structures. However, leveraging automatic differentiation and gradient descent, differentiable models can automatically learn to produce large amounts of parameters from geographic attributes through the embedded neural networks." Lines 393-404

"This study builds a benchmark and a basis for model selection and diagnosis for the next-generation global hydrologic modeling, which previously did not learn from such large observations. With rigorous tests at global scale, this study proves that differentiable models are strong candidates as global water models. With powerful spatial generalization ability, they can be applied to characterizing the hydrologic processes in ungauged regions by leveraging learned information in data-rich continents. Differentiable models in this study have already learned the generic and robust relationships between geographic attributes and physical model parameters from thousands of global catchments. Therefore, these models can be easily applied towards providing seamless global hydrologic modeling with parameters directly generated from worldwide geographic attributes. Future work can use such models to produce global hydrologic fluxes while enhancing some process representations in extremely arid, glaciated, or heavily human-influenced basins." Lines 427-435

Tsai, W. P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., ... & Shen, C. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. Nature communications, 12(1), 5988. doi: 10.1038/s41467-021-26107-z.

L200-205: Employing FHV and FLV is a thoughtful model evaluation strategy. In contrast to relying solely on integrated metrics like KGE, FHV and FLV offer a more thorough evaluation of model performance. Has the author considered whether integrated metrics such as KGE are suitable for capturing the distribution characteristics of state variables? Additionally, is the loss

function used during the neural network parameter determination suitable for the data's distribution characteristics? This is crucial, as KGE includes a term related to Pearson correlation coefficient, which may not be applicable to distributions beyond normal. Such considerations are essential to avoid potential misguidance in the model calibration process.

We used multiple metrics such as KGE, NSE, FLV and FHV to comprehensively evaluate model performance, but KGE was not the main objective function used to train the models in this study. The loss function was defined as a mixture of two parts based on the root-mean-square error (RMSE). We are aware of the distribution issue raised by the reviewer and thus employed this loss function. As shown in our previous study and the below equation (Feng et al., 2022), we applied a logarithmic transformation on the discharge data which made the residual more normal, and then further calculated RMSE on the transformed data as one part of the loss function. This part aims at improving the low flow representation. The second part is RMSE directly calculated on the discharge data, and we did a weighted combination of these two parts. We previously tested using MSE, RMSE, NSE as loss functions and they produced quite similar evaluation results, which implies the selection of calibration objectives was not as crucial as we initially thought. However, this weighted two-parts RMSE with data transformation applied can largely mitigate the issue of peak flow being over-emphasized from using the above-mentioned objective functions.

$$\widehat{Q}^{t} = Log_{10}(\sqrt{Q^{t}} + 0.1)$$

We added these statements in the text to clarify the loss function used to train the models:

"As in Feng et al., (2022), we employed the loss function based on root-mean-square error (RMSE) with two weighted parts. The first part calculates RMSE directly on the simulated and observed discharge, while the second part calculates RMSE on the transformed discharge records to improve low flow representations." Lines 179-182

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process‐based models with multiphysical outputs can approach state‐of‐the‐art hydrologic prediction accuracy. Water Resources Research, 58(10), e2022WR032404.

L208-212 & Figure 3: Are the evaluation results corresponding to the training set, the testing set, or the overall results from both? The assessment outcomes should be provided separately for the training set and the testing set to enable a clearer evaluation of the model's performance and identification of any issues within the model.

All the evaluation results reported in this manuscript are for the independent **testing dataset**. We completely agree that this is very important for fair model evaluation. We added and modified the following statements to clarify how we evaluate the model performance and report metrics.

"All the reported performance metrics in this study are from model evaluation on the testing dataset, which is not seen by the model during the training process." Lines 225-227

"For the temporal generalization experiment, the models were trained for the period of 2000 to 2016 on all global basins, and tested for the period of 1980 to 1997. Without spatially holding out any basin during training, this experiment aims at evaluating the model's generalizability in the time dimension by testing it in a different time period. The other two spatial generalization experiments serve as the true litmus tests for evaluating the effectiveness of regionalization schemes, i.e., how well the model can be applied to basins that have never been seen during training." Lines 203-208

L222-223: The structural issues might explain the errors with peak flow in differentiable models. However, it raises a question as to why LSTM also exhibits errors in peak flow. As mentioned earlier, the utilization of inappropriate evaluation metrics could contribute to the low FHV across all models. It requires a more in-depth consideration of how metrics impact the calibration results.

Our hypotheses for the causes of peak flow bias even in LSTM are (i) underestimation of precipitation inputs especially for large storms; (ii) to capture extremes, we need smaller basins and probably will need to run models on hourly time steps to capture the spatial heterogeneity in rainfall and nonlinear effects; (iii) extremes are not well observed so the training of neural networks (NNs) may be problematic for extremes; (iv) as you mentioned, some consideration with calibration metrics as well, but some previous efforts including from Kratzert et al., have studied this pretty extensively and they were not able to find a way to largely improve extreme metrics. We must mention that all these studies with different models and calibration metrics have shown the underestimation of peak flow (please see the FHV column in Table 3 in Kratzert et al., 2019), while integrating forcing data from several sources could moderately mitigate this underestimation (Kratzert et al., 2021), suggesting the forcings to be an important source for this issue. Therefore, the effects of the last hypothesis on the calibration metric could be limited, especially given the use of MSE, RMSE, NSE; these mean-square based metrics already emphasize large peak flow values. We think much more effort is needed to improve extreme representations and we look forward to community effort and collaboration in studying this topic.

We modified these statements in the text to discuss this point:

"However, for the peak flow predictions, the LSTM and differentiable models were quite similar, and they all underestimated the observed peaks (FHV in Figure 3). The underestimation for peak flows is consistent with what was found in previous studies. For example, all the physical and deep learning models have significant negative peak flow bias when benchmarked in the CONUS dataset (Feng et al., 2020; Kratzert et al., 2019b). We hypothesize that the systematic underestimation of peaks may be partially related to bias in precipitation forcings. MSWEP is based on the ERA5 reanalysis, which is known to underestimate precipitation peaks (Beck et al.,

2019). Furthermore, the use of basin-averaged, daily-averaged precipitation may further suppress the peaks (Chen et al., 2017). In addition, the errors with peak flow could also be partly due to some numerical and structural issues with the differentiable models, e.g., numerical errors introduced by the explicit and sequential solution scheme of HBV with excessive use of threshold functions that lead to different results when the sequence changes, and structure limitations, e.g., deeper groundwater storage cannot feed back to the upper layers. Given the commonality of this issue, we call for community efforts and collaboration to address this issue." Lines 242-252

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences, 23(12), 5089-5110.

Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. Hydrology and Earth System Sciences, 25(5), 2685-2703.